Geophysical Journal International

Geophys. J. Int. (2014) **197,** 458–463
Advance Access publication 2014 January 20

doi: 10.1093/gji/ggt516

An earthquake detection algorithm with pseudo-probabilities of multiple indicators

Z.E. Ross and Y. Ben-Zion

Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089-0740*, USA. E-mail: zross@usc.edu*

Accepted 2013 December 19. Received 2013 December 18; in original form 2013 November 15

SUMMARY

We develop an automatic earthquake detection algorithm combining information from numerous indicator variables in a non-parametric framework. The method is shown to perform well with multiple ratios of moving short- and long-time averages having ranges of time intervals and frequency bands. The results from each indicator are transformed to a pseudo-probability time-series (PPTS) in the range [0, 1]. The various PPTS of the different indicators are multiplied to form a single joint PPTS that is used for detections. Since all information is combined, redundancy among the different indicators produces robust peaks in the output. This allows the trigger threshold applied to the joint PPTS to be significantly lower than for any one detector, leading to substantially more detected earthquakes. Application of the algorithm to a small data set recorded during a 7-d window by 13 stations near the San Jacinto fault zone detects 3.13 times as many earthquakes as listed in the Southern California Seismic Network catalogue. The method provides a convenient statistical platform for including other indicators, and may utilize different sets of indicators to detect other information such as specific seismic phases or tremor.

Key words: Time-series analysis; Probability distributions; Earthquake dynamics; Earthquake source observations; Dynamics and mechanics of faulting.

1 INTRODUCTION

Automated algorithms for earthquake detection require the use of indicator variables that tend to correlate in amplitude with the arrival of earthquake waveforms. These indicators can operate in either the time or frequency domain. One common example is the ratio (e.g. Allen 1978) of a short-term moving average (STA) to a long-term moving average (LTA). Polarization analysis using singular value decompositions (e.g. Rosenberger 2010) or covariance matrices (e.g. Jurkevics 1988) can provide additional indicator variables for detections. Kurzon *et al*. (2014) have used polarization analysis to separate real-time data into channels associated with *P* and *S* waves, and analyse various indicator variables to achieve significantly improved *P* and *S*-wave picks. Withers *et al.* (1998) compared many different indicator variables and find that the STA/LTA is often the most reliable. At present seismometers commonly record in a continuous mode, providing new opportunities for detecting many small events. An ideal detection framework will identify all events recorded by a network down to the noise level. However, the current methods fall considerably short of this goal, and specialized techniques such as using templates (e.g. Peng & Zhao 2009; Yang *et al.* 2009) often detect many more small events than the standard automated methods.

To improve the performance of automated algorithms, multiple STA/LTA detectors can be used with different STA and LTA windows and/or different frequency bands. Each STA/LTA indicator can be evaluated independently with different trigger thresholds. If these thresholds are met simultaneously by a sufficient number of indicators across enough stations (in the association phase of the algorithm), a detection can be made. The primary issue associated with this approach is that any one indicator may often trigger erroneously when an earthquake is not present. The cut-off threshold for triggering is typically adjusted so that a suitable number of false detections are allowed (e.g. Nippress *et al.* 2010), and the smallest detected earthquakes are therefore limited by the accuracy of a single indicator. As such, the collective amount of information available from running multiple STA/LTA detectors is not used in a way that takes full advantage of the redundancy.

In this study, we describe a method that circumvents this problem by combining the information from each indicator together before any thresholds are applied to the data. This is done in a non-parametric way that does not involve weighting schemes and offers great flexibility in choosing which indicator variables to use. The primary advantage is that the joint detection threshold used on the combined set of information can be lowered significantly compared to any single indicator. This results in more detections at a more

Table 1. Best performing set of 10 different STA/LTA detectors in the 24-hr test period.

STA (s)	LTA (s)	High pass (Hz)
3	10	3
3	15	3
3	20	3
3	25	3
3	30	3
2	5	5
2	7	5
2	9	5
2	11	5
2	13	5

reliable rate. We demonstrate the potential of the algorithm in a small region around the San Jacinto fault zone, and compare the results to those obtained by the Southern California Seismic Network (SCSN).

2 METHODOLOGY

Our methodology for automated detection consists of several simple steps. We first choose the type and number of indicators to be used in the detection process. We focus on STA/LTA indicators with 10 different short- and long-term windows and different frequency bands (Table 1). The specific employed indicators are chosen to detect small earthquakes by a local network. In this section, we describe the approach used after testing the methodology; the details of the tests are documented in the subsequent section. Continuous waveforms are separated into 1-hr segments that overlap slightly to ensure that no events are missed at the edges. We find that 1 hr of data is long enough to characterize the wavefield appropriately relative to the timescales of $M < 4$ earthquakes, and is computationally efficient. Also, 1 hr is short enough to retain the unique temporal characteristics of the noise background for that interval. We assume that the majority of the signals in 1 hr of data are associated with ambient noise rather than earthquakes.

Fig. 1(a) shows an example of a 1-hr vertical component trace on 2013 March 27 containing 21 earthquakes. Nearly all of these earthquakes have $M < 2.0$ and only several are visible with the scale of the figure. The 10 different STA/LTA detectors are used on the vector magnitude (i.e. $\sqrt{N^2 + E^2 + Z^2}$) of each hour window (Fig. 1b). Similar results are obtained if the individual components are used and the requirement of N stations in the association phase is replaced with $3N$ traces. The result of a single STA/LTA detector on the vector magnitude for this time window is shown in Fig. 1(c), using an STA window of 3 s, LTA window of 10 s and a high-pass filter at 3 Hz. Large values of the STA/LTA ratio indicate significant changes in the amplitudes of the wavefield over time, while accounting for variability in the noise level. We now consider only the values of STA/LTA during this time window while temporarily discarding the time associated with each value. This leads to a sampling distribution (Fig. 1d) that characterizes the range and likelihood of different STA/LTA ratios that were observed during that hour.

Since we employ here a total of 10 different STA/LTA detectors, we obtain 10 different sampling distributions. For each detector, we use its distribution as a reference and then compare each value of the STA/LTA time-series to its relative ordering within the sample. To do this, each STA/LTA distribution is used to compute an empirical cumulative distribution function (ECDF; e.g. D'Agostino & Stephens 1986). The ECDF describes the cumulative fraction of values less than or equal to a given number

$$F(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq s\}, \tag{1a}$$

where F is the value of the ECDF at a given value of STA/LTA (denoted by s), x_i is the ith value of the particular sample with n total values, and

$$\mathbf{1}\{x_i \leq s\} = \begin{cases} 1 & \text{if } x_i \leq s, \\ 0 & \text{if } x_i > s. \end{cases} \tag{1b}$$

Fig. 1(e) shows the ECDF corresponding to the distribution in Fig. 1(d). The horizontal axis indicates the STA/LTA size and the vertical axis indicates the fraction of values in the 1-hr window that are less than or equal to the corresponding STA/LTA amount. The lowest STA/LTA ratios on this plot have ECDF values close to zero, whereas the largest ones have ECDF values close to one.

We use the ECDF to define a mapping, through which we transform each STA/LTA value, s, in the time-series (e.g. Fig. 1c) with the corresponding $F(s)$. This converts the entire time-series into a new one having a range of [0, 1], with each value exactly equal to the original value's percentile relative to all others. We call this a pseudo-probability time-series (PPTS), because at each time step we take the corresponding probability value from the ECDF. The 'pseudo' distinction signifies that we use the PPTS solely to describe relative ordering within a sample. The PPTS corresponding to the example time-series (Fig. 1b) is shown in Fig. 1(f). It is clearly more erratic than the STA/LTA series in Fig. 1(c) and regularly reaches values greater than 0.9 when no earthquakes are present. Since by definition no more than 10 per cent of the values are allowed to reach 0.9, the mapping becomes very useful when used with multiple indicators.

In the present application, we have 10 PPTS for each 1-hr window with values in the range [0, 1], which can be combined by simple multiplication at each time step. While a single PPTS frequently reaches high values over the course of the time window (Fig. 1f), the likelihood of all 10 PPTS having large values is extremely low. Furthermore, low to medium values tend to cancel each other, leading to clear robust peaks in the product PPTS (Fig. 1g). Instead of a set of 10 STA/LTA detectors that are used with 10 *ad hoc* detection thresholds, we have one joint PPTS detector with a single detection threshold. We can compute this for each station just like a standard STA/LTA detector, but the advantage is that with our method, the redundancy built into the product PPTS allows the single cut-off threshold to be lowered more than any one STA/LTA detector.

It may seem that the product PPTS in Fig. 1(g) is much noisier than the basic STA/LTA series in Fig. 1(c), but there are distinct advantages to the former. First, the PPTS is bounded between [0, 1] whereas the STA/LTA is only required to be greater than or equal to zero. The bounds imposed by the PPTS, along with the fact that it is derived from multiplication, make it progressively more difficult to attain high values. As an example, when all 10 PPTS have a value of 0.99 or higher the product is ∼0.90, while if all 10 have a value of 0.93 or lower the product is ≤0.49. If any one PPTS has a value of 0.5 (the median) at a given moment, the product will automatically be less than or equal to this. We find that when the product PPTS value reaches ∼0.8, there is essentially always an earthquake or strong departure from the preceding noise background that should be recognized for that station. The existence of this feature at other stations can be used to refine the detection further. However, we

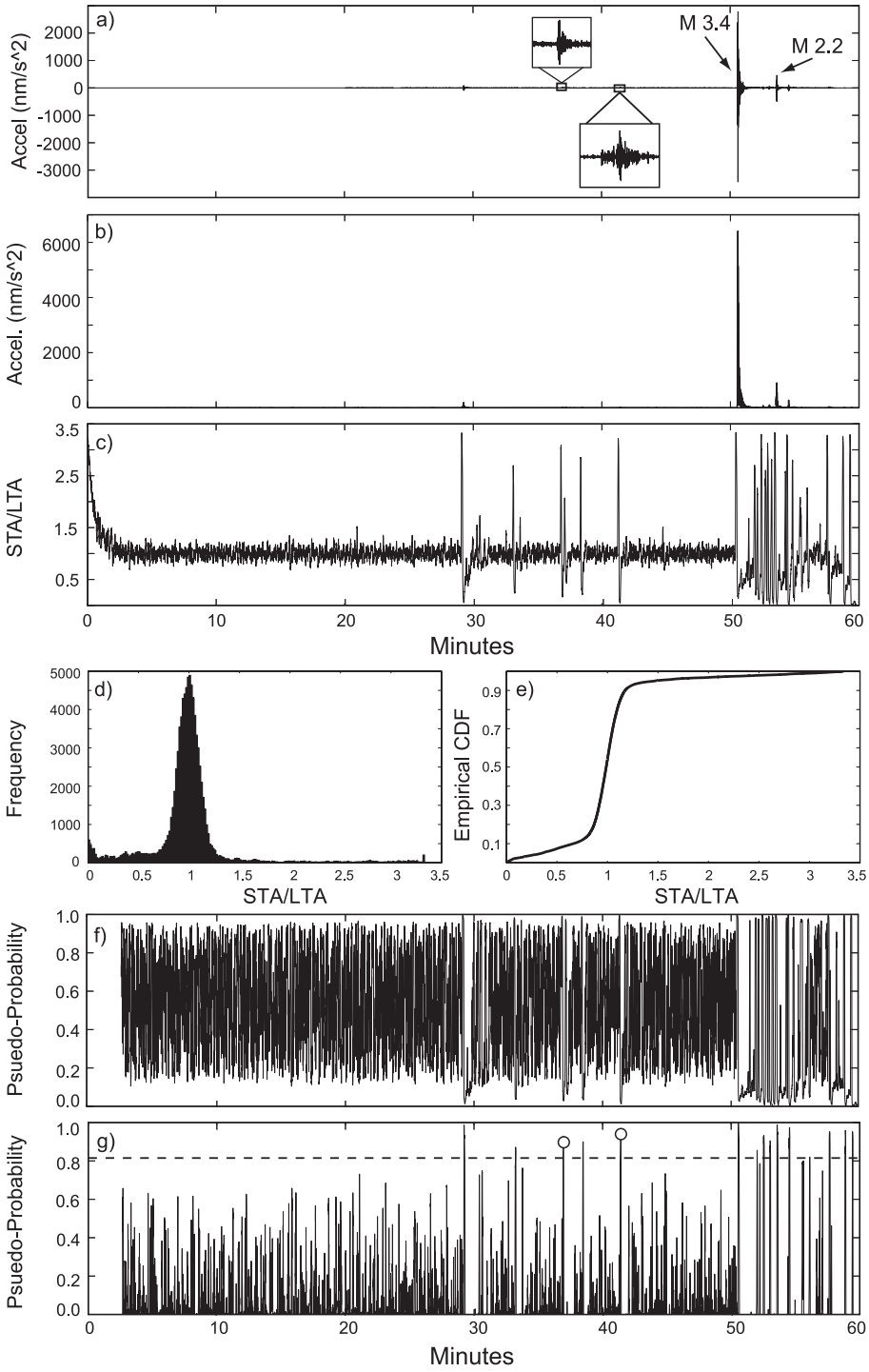

Figure 1. A visual illustration of the detection algorithm. (a) Example 1-hr vertical seismogram from the orange station in Fig. 2 with 21 earthquakes. Two example events detected are shown with an amplified scale. (b) Vector magnitude of the three components for the same 1-hr data. (c) STA/LTA detector trace with an STA window of 3 s and LTA window of 10 s. (d) Histogram of the obtained STA/LTA values. (e) Empirical CDF of the STA/LTA values, giving the percentile of each value relative to all others in that hour. (f) A pseudo-probability time-series for the examined hour of data. (g) A joint PPTS obtained by multiplying 10 individual PPTS at each time step. A probability threshold of 0.8 (dashed line) is generally found to distinguish a signal from noise at a given station. There are 17 events above the threshold with the missed four buried in the coda of the largest ones. Circles indicate the peak values for the two events magnified in (a).

find that the product PPTS even at a single station performs well in flagging out unusual portions of the data.

In the following, we adopt a very simple association scheme using the product PPTS detectors for testing the methodology. For each product PPTS, we initiate a trigger window when a value of 0.3 in probability units is reached, and the window is turned off when this falls to 0.1. We require that the window be a minimum of 2 s in duration and reach a peak value of at least 0.82. If these criteria are not met the window is discarded. From the set of all remaining trigger windows for each station, we define detections

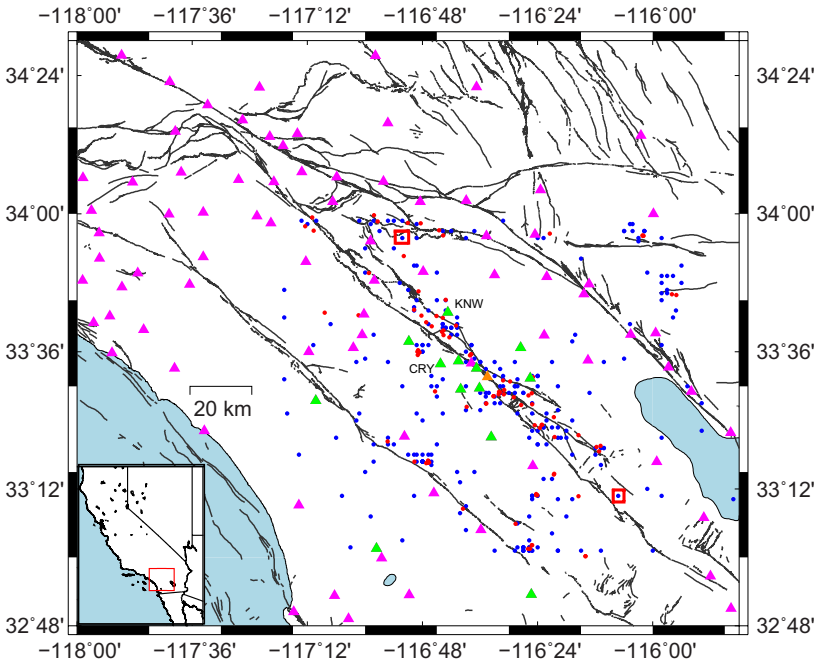

Figure 2. Events detected by the developed method (blue dots) during 7 d of data recorded by the 13 ANZA stations (green triangles) that form a subset of the SCSN stations (purple triangles). The algorithm detects 3.13 times the number of earthquakes listed in the SCSN catalogue for the examined area and time (red dots). The red boxes mark the locations of two events shown in Fig. 3.

when a minimum of six stations have overlapping trigger windows (no minimal overlap time is required). These detection criteria are found to have a false detection rate of ∼2 per cent based on the examples presented in Section 4.

3 INDICATOR VARIABLE TESTING

Here we discuss the tests leading to the best parameter combination for the detectors used in this work, and the choice of the number of indicators. We performed an exploration of the STA, LTA and high-pass filter parameter space for 10 detectors with a 24-hr window of time, starting on 2013 January 1 00:00:00. The employed recordings are from 13 ANZA stations around the San Jacinto fault zone (green triangles in Fig. 2). We initially focused on finding parameters that detected all the 19 local earthquakes listed by the SCSN (available at http://www.scecdc.scec.org/) in the considered region of space and time. This was achieved using first a coarse exploration of the parameter space, followed by more detailed exploration in the neighbourhoods of well-performing parameters, using many combinations including those listed in Table S1 (Supporting Information). The best performance was obtained by the 10 parameters of Table 1, leading to detection of 82 earthquakes in this 24-hr window. Of these, 81 events were visually confirmed at six or more stations, with only one false detection for that combination. For comparison, using only one of the 10 detectors with STA and LTA windows of 3 and 15 s, and the same association scheme with on and off trigger thresholds of 3.5 and 1, detected just 21 earthquakes which is close to the SCSN results.

Using the parameters of Table 1, we tested the number of indicators necessary to achieve stability in the product PPTS for the first hour of the data previously examined. The full 24 hr was not used because a large number of false detections are present when using only a few indicators. We started with two indicators, and ran the algorithm on each station. We recorded the number of earthquakes detected and each was visually confirmed. It was also noted how many earthquakes were missed. The results of repeating this process while progressively increasing the number of indicators are summarized in Table 2. The ordering of the indicators was found to be relatively unimportant as each indicator is an STA/LTA detector. It can be seen that it takes eight indicators of the form used to achieve a stable result. We have, however, used 10 consistently throughout this study because we find that the extra stability does not affect the performance of the algorithm.

Table 2. Results of testing the number of indicators needed for PPTS stabilization.

No. indicators	No. detected	Per cent true	Per cent false	No. missed
2	23	4	96	8
3	53	9	91	4
4	35	17	83	3
5	22	27	73	3
6	16	50	50	1
7	14	57	43	1
8	9	100	0	0
9	9	100	0	0
10	9	100	0	0

4 ILLUSTRATION FROM THE SAN JACINTO FAULT ZONE REGION

With the parameters and number of indicator variables optimized, we now demonstrate the algorithm in comparison to the SCSN. The region of study is the Southern California plate boundary area depicted in Fig. 2. As before, we use only the 13 ANZA stations shown. We focus on a 7-d window of time beginning on 2013 January 1 00:00:00, in which the SCSN detected 91 earthquakes (red dots in Fig. 2). The stations used by the SCSN include the green triangles as well as all the purple ones. The SCSN uses more complex event association than we do, but their detection criteria requires

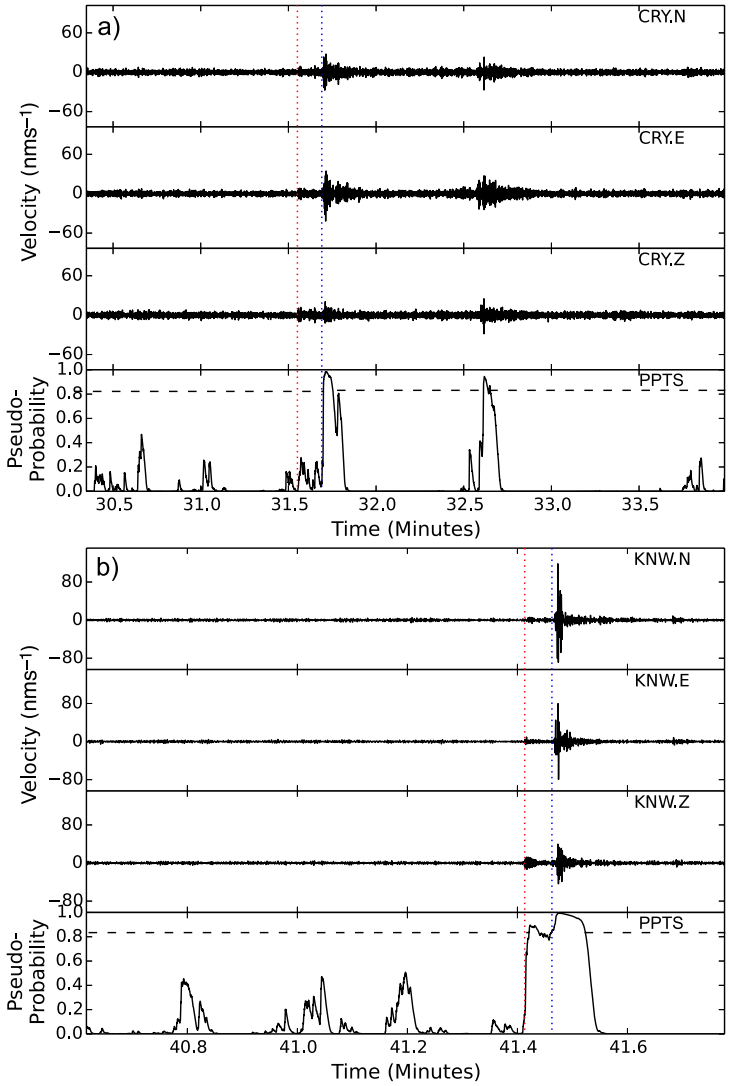

Figure 3. Seismograms of two example detected earthquakes and corresponding product PPTS. (a) Three-component traces and joint PPTS (at station CRY) for the event in the SE red box of Fig. 2. The red and blue dotted lines indicate, respectively, the *P* and *S* arrivals of the earthquake detected by a minimum of six stations. The black horizontal dashed line indicates the detection threshold. (b) Corresponding results for the event in the NW red box of Fig. 2 at station KNW.

essentially five or more high-quality phase picks (E. Hauksson, personal communication, 2013), which is similar to our approach. We ran our algorithm on the vector magnitude of the three-component traces for each station using the detection criteria described in Section 2 and the 10 STA/LTA detectors listed in Table 1. Each detection was visually confirmed, and *P* and *S*-wave arrival times were picked by hand.

The detection algorithm identified 456 possible earthquakes, of which 20 did not have at least six clear *P* and *S* picks. Of the remaining earthquakes, 419 were located successfully using the HYPOINVERSE code (Klein 2003) with the velocity model of Hadley & Kanamori (1977), and a total of 285 were found to be in the region of Fig. 2 (blue dots). Three-component seismograms for two of these earthquakes are shown in Fig. 3 along with the example PPTS at individual stations used to make the detection. The detected earthquakes are indicated by red and blue lines and are associated with a clear peak of the PPTS. In Fig. 3(a), the algorithm identified only the *S* wave, while in Fig. 3(b) with higher signal-to-noise ratio both *P* and *S* waves are detected. The *S*-wave picks in Fig. 3 were obtained using the peak PPTS value as a starting position, while the *P*-wave picks were obtained using a lower PPTS trigger threshold of 0.3. Both picks were refined using nearby peaks of the time derivative of PPTS values (Fig. S1) derived from the individual trace components (vertical for *P* and horizontal for *S*). These results were found to be in good agreement with phase picks based on the Rosenberger (2010) algorithm. There are additional transient signals visible in Fig. 3 that were not detected at six or more stations. Some of these are likely to be earthquakes and may be detectable by reducing the number of stations with corresponding signals to a smaller (and spatially more compact) subset and/or using other indicators. However, our present purpose is to demonstrate the utility of the method and we leave efforts to detect additional smaller events to future work. Most of the newly detected events are concentrated in areas where there was already some seismicity detected by the SCSN, suggesting earthquake sequences with considerably more events than known before. This significant increase in the number of events was obtained using only 13 of the 41 SCSN stations in this region.

5 DISCUSSION

Our detection algorithm combines statistically normalized information from multiple indicators to achieve a more robust indicator variable. It offers considerable flexibility in terms of which indicators to employ and what parameters to use them with. In this study, we used exclusively STA/LTA detectors, but other indicator variables that produce a time-series can also be used. These include polarization information such as pseudo-incidence angle (e.g. Rosenberger 2010), or the various types of amplitude-based detectors summarized by Withers *et al.* (1998). While we focused on local earthquake detection in the San Jacinto fault zone area, other users may select indicator variables that are more relevant to different regions and scales, or signals. Overlapping events present a challenge for the PPTS and may result in a single continuous signal, but some earthquakes buried in coda are identifiable. The method can work with larger earthquakes but requires appropriately tuned parameters.

The algorithm is based on exploiting the redundancy in the peaks of a set of indicator variables at common time points, which should occur only for unusual portions of the seismograms. With enough indicator variables, the likelihood of them all producing large values is low. This can be used to produce a time-series that has much more robust peaks than those from any one indicator variable. The core of the method uses multiplication of a set of PPTS at each time step, each varying in the range [0, 1]. This common normalized range facilitates separation of the background noise from earthquake arrivals by multiplication of the individual PPTS. We also tested adding the PPTS, but this does not yield clear results even with various weighting schemes (which we try to avoid). The idea of multiplication was motivated by the set theory intersection operation, which describes the likelihood of a set of outcomes being true simultaneously. We use the term pseudo-probability, because there is an obvious correlation neglected between these indicators as they all measure a quantity derived from the same time-series. This is not a problem here because we are not interested in actual probabilities or forecasting, but rather a quantity that measures the simultaneous coherence of a set of indicators.

The choice of running the algorithm on 1-hr intervals is due to several reasons. The first is that the ambient noise changes on many timescales, and we are interested in a window of time that is representative of the noise around the time of earthquake arrivals. The algorithm employs 1-hr intervals to obtain a sample of STA/LTA values, which are used to characterize the state of the wavefield over that hour via the ECDF. Each value of the original STA/LTA time-series is then transformed into its percentile relative to all other values. By using the window of time surrounding each value, we get a relevant reference distribution that is representative of each examined time. If the 1-hr window is extended significantly, it may ultimately lead to increased sensitivity in the PPTS because some of the values generating the ECDF may be too different than those observed at a given moment. A time window that is too narrow could lead to less sensitive results because the ECDF may be biased towards the most recent values. The choice of 1 hr seems to be a good compromise, but changing the duration around that value has little effects on the results.

As demonstrated in Section 4, using the technique with a simple detection scheme on a small network near the SJFZ detects over three times as many earthquakes as listed in the SCSN catalogue. The method may perform even better by combining different types of indicators, instead of just STA/LTA detectors with varying parameters. Phase picking is easier if an earthquake has already been detected near a particular point in time. The PPTS of individual components and their time derivatives can be useful for automatic picking of *P* and *S* phases because the phases generally arrive at the same relative location within the PPTS peak for each event. Improving earthquake detection algorithms is essential for utilizing the full potential of recorded data. Reliable detections of smaller events will lower the completeness magnitude of seismicity and focal mechanism catalogues, and may reveal key new information on the physics of faulting. Adaptation of the algorithm for real-time detection may be possible with recursive updating techniques.

ACKNOWLEDGEMENTS

The study was supported by the National Science Foundation (grant EAR-0908903). The manuscript benefitted from constructive comments of two anonymous referees.

REFERENCES

Allen, R., 1978. Automatic earthquake recognition and timing from single traces, *Bull. seism. Soc. Am.,* **68,** 1521–1532.

D'Agostino, R.B. & Stephens, M.A., 1986. *Goodness-of-Fit Techniques,* Vol. **68,** CRC Press.

Hadley, D. & Kanamori, H., 1977. Seismic structure of the transverse ranges, California, *Bull. geol. Soc. Am.,* **88,** 1469–1478.

Jurkevics, A., 1988. Polarization analysis of three-component array data, *Bull. seism. Soc. Am.,* **78,** 1725–1743.

Klein, F.W., 2003. 85.8 The HYPOINVERSE2000 earthquake location program, *Int. Geophys.,* **81,** 1619–1620.

Kurzon, I., Vernon, F.L., Rosenberger, A. & Ben-Zion, Y., 2014. Real-time automatic detectors of *P* and *S* waves using singular value decomposition, *Bull. seism. Soc. Am.,* in review.

Nippress, S.E.J., Rietbrock, A. & Heath, A.E., 2010. Optimized automatic pickers: application to the ANCORP data set, *Geophys. J. Int.,* **181,** 911–925.

Peng, Z. & Zhao, P., 2009. Migration of early aftershocks following the 2004 Parkfield earthquake, *Nature Geosci.,* **2,** 877–881.

Rosenberger, A., 2010. Real-time ground-motion analysis: distinguishing *P* and *S* arrivals in a noisy environment, *Bull. seism. Soc. Am.,* **100,** 1252–1262.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S. & Trujillo, J., 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection, *Bull. seism. Soc. Am.,* **88,** 95–106.

Yang, H., Zhu, L. & Chu, R., 2009. Fault-plane determination of the 18 April 2008 Mount Carmel, Illinois, earthquake by detecting and relocating aftershocks, *Bull. seism. Soc. Am.,* **99,** 3413–3420.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Three-component seismograms of the event in Fig. 3(b) and PPTS calculated from individual components instead of the vector magnitude.

Table S1. Results of tests for well-performing detection parameters using 10 different STA/LTA detectors (http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggt516/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.